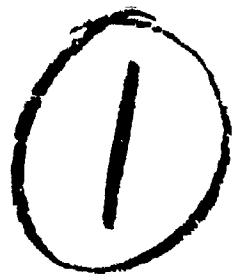

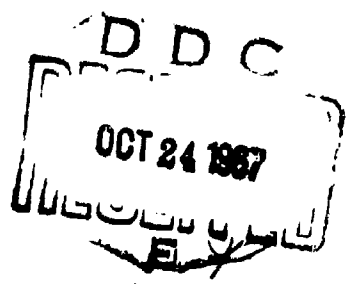

UNIVERSITY of PENNSYLVANIA

The Moore School of Electrical Engineering

PHILADELPHIA, PENNSYLVANIA 19104

19

University of Pennsylvania
THE MOORE SCHOOL OF ELECTRICAL ENGINEERING
Philadelphia, Pennsylvania

WORD ASSOCIATION EXPERIMENTS --

BASIC CONSIDERATIONS

by

Don Stone

August 1966

The Moore School Information Systems Laboratory

Morris Rubinoff
Principal Investigator

S. Bergman
H. Cautin
T. Johnson
F. Franks
T. C. Lowe
J. Lucas
S. Newman
F. Rapp
E. R. Rubinoff
D. Stone

AF49(638)-1421

University of Pennsylvania
THE MOORE SCHOOL OF ELECTRICAL ENGINEERING

TO: Project VECTOR/ROSE

FROM: Don Stone

DATE: 5 August 1966

SUBJECT: Word Association Experiments -- Basic Considerations

I. INTRODUCTION

The process of indexing incoming documents (generally by assigning terms, i.e., words or word phrases, to them to indicate their content) is common to most retrieval systems. Classification or assignment of such terms or documents to classes, is likewise an important operation in such systems. For example, a very rudimentary classification system might consist of classes of the form "all documents indexed by the term X".

There are various devices which supplement the basic process of indexing and thereby increase retrieval effectiveness. These are summarized by Cleverdon[1], who is in the process of testing their comparative utility. Salton's SMART[13] system is also designed to enable testing of such devices. Doyle[2] has pointed out two somewhat contrasting approaches to information retrieval, each one involving a device or a group of devices for generating additional information to aid in retrieval.

The first approach is to concentrate on the classification of documents or terms, perhaps using a hierarchy (which might be automatically generated) to exhibit various relations among them. The other approach builds in the concept of coordinate indexing, which permits requests for documents indexed by logical combinations of terms. The extensions to

coordinate indexing which form this second approach are probabilistic indexing and statistical association techniques.

Probabilistic indexing was proposed in 1960 by Maron and Kuhns[3]. The basic idea is that in the process of manually indexing a document, the indexer associates with each document a set of ordered pairs rather than a set of terms. The first element of each ordered pair is a term and the second a subjective estimate of the relevance of the term to the document, or more precisely, the probability that the document will be considered relevant to a request containing that term. This assignment permits the computation of a relevance score of a document relative to a request.

In the use of statistical association techniques, a coefficient of association is computed for each pair of term, indicating the degree to which the two terms co-occur in the document collection. This co-occurrence can be either in the list of terms for a particular document (the document's index set) or by association in the document text, say, by appearing within a certain distance of each other. This tendency of certain terms to co-occur provides a relationship which can be exploited in retrieval: a search can be made not only for documents indexed by the requested terms, but also for documents indexed by terms closely associated to the specified ones[9]. As a result, documents may be retrieved which are useful to the searcher even though indexed by a combination of terms he would not have thought of suggesting.

Historically, adherents of the two approaches have often formed opposing camps. Several years ago, there was much emphasis among computer-oriented people on the coordinate indexing approach as a solution to deficiencies of classification systems such as their difficulty in handling the coupling between interdisciplinary fields. More recently,

the assistance which a classification system can provide a requester has led to research on automatic classification. Statistical association of terms has served to some extent as a tie between the two approaches, since in addition to its use in the first area (in expanding requests to include associated terms), it is used in many schemes for the automatic generation of classification systems. Doyle believes that a synthesis of the two approaches will provide particularly satisfactory results. It is interesting to note that Cleverdon's research has led him to the opinion that for a well indexed set of documents, most of the various devices or approaches are potentially capable of equally good performance, and moreover, the basic process of indexing is not crucial to good retrieval.

The purpose of this document is to review prior research in these two areas, relate it to Project VECTOR/ROSE, and to recommend experiments for determining the extent to which word association techniques can be introduced into the Moore School Information System.

1.. INDEXING AND VOCABULARY GENERATION CONSIDERATIONS

Before going into details on the two approaches to information retrieval, both of which start with indexed documents, it might be beneficial to survey the research in automatic indexing. The classic article on this topic is that of Edmundson and Wyllys[4], in which they suggest that the index terms for a document could be the terms whose relative frequency in the document is significantly higher than their relative frequency in the literature as a whole. Experimental results reported by Damerau[5] are encouraging. O'Connor[6] organized an experiment in the automatic assigning of two medical terms to documents in which they do not necessarily appear. The approach is somewhat analogous to searching for dpcuments to supply information on these terms in an information retrieval system incorporating statistical associations, a thesaurus, and various other aids. O'Connor's survey article[7] discusses various operations (such as indication of antecedents of pronouns) that might be done on a document prior to the actual process in which index terms are assigned, and gives the impression that adequate completely automatic indexing is some distance in the future.

A closely allied topic which should be mentioned at this point is automatic vocabulary generation. Here the object is to find the words which are important in the document collection as a whole rather than the important words in a single document. A method for doing this is treated by Dennis[8]. The hypothesis is that the non-informative words will have roughly the same relative frequency within all the documents, but that the informative words will have a skewed distribution of relative frequency, i.e., in some documents they will appear with a comparatively

high frequency but in most documents they will have a very low frequency or not appear at all. Note the relation between this use of relative frequency and the approach of Edmundson and Wyllys. The hypothesis was tested by Dennis on a set of legal literature and the measure of skewness proved to be a better test of importance than several other measures.

III. STATISTICAL ASSOCIATION TECHNIQUES

Of the two approaches to information retrieval which this report analyzes, one (automatic classification) has no single pioneer paper. For the other (statistical association techniques, including what Doyle calls associative machine searching), the 1960 paper by Maron and Kuhns[3] is a landmark. Their concept of probabilistic indexing was discussed briefly earlier. This same paper also introduces the idea of statistical association of index terms. It suggests that if two terms "X" and "Y" appear together more frequently in the set of terms indexing a document than they would by chance, then a request containing "X" could be expanded to "X" OR "Y", thus improving recall (the fraction of relevant documents retrieved). Suppose someone is interested in documents about "cartoons". If the closely associated terms "animation", "Walt Disney", etc., are added to the request perhaps automatically, relevant documents which did not happen to be indexed under "cartoons" might be retrieved. The computation of a relevance score for documents with respect to a request permits discarding of documents with a low score and hence allows some control over precision (the fraction of retrieved documents which are relevant). The documents which have been discarded can be retrieved, if desired, in order of decreasing relevance score.

Another important early paper on statistical association techniques is that of Stiles[9]. He also makes association of a pair of terms depend on their co-occurrence in index sets of documents, and calls his measure of the association of two terms their "association factor". He points out that a pair of synonyms may have a fairly low

association factor, since they will not tend to be used to index the same documents. But on the other hand, they are both likely to have high associations with related terms. For example, "movie" and "motion picture" might have a low association factor, but they would both be highly associated with "production", "film", "theater", etc. Stiles would say "movie" and "motion picture" have a high second generation association. The system proposed by Stiles expands requests to include first and second generation associated terms, and computes document relevance scores based on the association factors between the terms of the document's index set and the terms of the expanded request.

In later papers Stiles advocates having the requester assign weights to the terms he specified, then presenting him with a list of associated terms and permitting him to add to his original list or to revise weights. Experimental results show that this approach can provide greater recall than pure coordinate indexing, and (as in Maron and Kuhn's system) the user can examine the retrieved documents in order of their estimated relevance to his request.

Doyle[10,11] proposed the concept of an association map, a two-dimensional display in which highly associated terms would be close to each other and would be joined by a line. The patterns of associations presented by such a map could suggest additional terms that a requester might add to his request, or could aid in the manual compilation of a thesaurus. Doyle later came to believe that a hierarchical association map, in which term hierarchies based on relative frequency of occurrence are displayed more prominently than pure statistical association, would be more effective than a map which

displayed only statistical associations. Most recently his attention has been centered on a different scheme for automatic hierarchy generation, discussed in the next section.

A detailed treatment of the mathematics of associative retrieval came from Guiliano and Jones in 1963 [12]. They propose a model in which a linear transformation of a request vector results in a response vector. The components of the request vector are the weights assigned to the different terms by the requester, and the components of the response vector are relevance scores for the documents in the system. The request vector is first premultiplied by an index term association matrix, resulting in what can be considered a modified request vector: the components corresponding to terms not in the original request but highly associated to those terms will now have a non-zero value. This vector will then be multiplied by a term-document connection matrix, whose typical element c_{ij} could be the weight assigned to term j as it indexes document i, obtained by probabilistic indexing. The resulting vector will be in the form of a response vector. Finally, an optional multiplication by a document association matrix could adjust the components (relevance scores) to take into account associations between documents. Various methods of obtaining the matrices associated with the transformation are discussed, and an electrical network analog for accomplishing the transformation is explained. In the latter, the components of the request vector are supplied to the network as currents, and the response components are produced as voltages.

The use of statistical association techniques is one of the options of the SMART Retrieval System developed by Salton at Harvard[13]. Association coefficients can be computed using either co-occurrence of a pair of terms or concepts in a sentence or co-occurrence in the index set of a document. Salton calls a group of closely associated terms a cluster. If one of the terms in a cluster is part of a request, the user may, if he wishes, add the other terms of the cluster to the request.

Several reports on recent research in the area of statistical association techniques as well as automatic classification are contained in a survey by Mary E. Stevens[8]. Discussion of possible applications of statistical association techniques to Project VECTOR/ROSE will be held until Section V.

IV. AUTOMATIC CLASSIFICATION

Two types of classification should be distinguished: classification of concepts or terms and classification of documents. A classification system for concepts is often a set of successive partitions of the knowledge to be classified. These successive partitions can be viewed as a tree structure where the top node (the root) represents the whole field of knowledge, the nodes one level down represent the large subdivisions of the field, each set of nodes on the third level represents a partition of a large subdivision, etc. The relation of a node to its successors is that of generic concept to more specific concept. Generally the nodes are labeled and all concepts or terms can be found as the label of some node in the hierarchy. One of the criticisms of classification systems is that it is often difficult to assign just one predecessor node, for example, if an interdisciplinary concept is involved. Due to problems such as this, several classification schemes permit a concept to be an immediate member of more than one class, i.e., they allow a node to have more than one immediate predecessor.

Classification of documents is a somewhat different idea, though often a concept classification system is adapted and a document is assigned to the class corresponding to its main concept. However, documents could be grouped according to the similarity of their index sets, hence avoiding explicit use of a concept classification.

One of the purposes of classification is to aid in assigning and identifying the physical location of documents in storage. But it also has another use: if a concept hierarchy is available, a user

can locate the terms he was thinking of putting in his request, and upon inspection of more general or more specific terms or parallel terms on the tree, he may alter his request to improve his chances of retrieving what he wants. Similarly, if there is a document classification scheme and the user knows one document that is relevant to his needs, he may want to inspect all the other documents in the same class.

Automatic classification schemes fall into two categories: automatic placement of documents or terms in a predetermined classification system, or automatic generation of the classification system and placement of documents or terms in it. Most of the systems take indexed documents as input and make use of the statistical association of terms or some related concept. The majority of the systems work with a two-level hierarchy, corresponding to a single partition.

Another way of looking at classification systems is via vector spaces. Suppose all knowledge is considered to be an m-dimensional vector space. Then one might represent concepts as points in this space, and similar concepts would be closer together than unrelated ones. The successive partitions could be visualized as sets of hypersurfaces in the space. (Material relevant to this viewpoint is contained in some of the literature on pattern recognition.) Some of the people who view classification in this way are not very explicit about what the axes would be or how points are located in the space, e.g., Maron and Kuhns, and Hayes[14]. Borko's factor analysis approach results in a space of smaller dimension than the original in which concepts are axes rather than points, and documents are represented as points. A request can be viewed in this framework as a point or set of points in the vector space, and

the retrieval problem can be characterized as finding document points near to the request point or points by some distance measure. Maron and Kuhns consider the request space and document space to be distinct and retrieval to be a mapping from the former to the latter.

Maron[15] proposed one of the earliest methods for automatic classification of documents, utilizing the occurrence of key words in abstracts and Bayesian probability formulas. He used a predetermined classification system, as did Williams. As an extension of Edmundson and Wylly's idea that each field of knowledge is characterized by a special distribution of word frequencies, Williams[16] experimented with the use of multiple discriminant functions in the automatic classification of documents. This procedure seemed to be quite successful in classifying solid state physics abstracts.

One of the earliest approaches to automatic classification of terms, and one which involved automatic generation of the categories, was pursued by Needham, Parker-Rhodes, et al. at the Cambridge Language Research Unit in England[17,18]. Their system forms non-exclusive subsets (called clumps) of a set of terms, such that terms in a given subset are more closely related to each other than to terms in other subsets. Their idea is that documents can be indexed by clumps rather than individual terms. Another approach, due to Borko and Bernick[19], is to use factor analysis to group highly correlated terms into factors or categories. Documents are automatically assigned to one of the categories derived. Latent class analysis, proposed by Baker[20], is another scheme for reducing the large set of terms to a smaller set of classes. This process utilizes correlations of not only pairs of terms, but also

triples, or even larger sets, if desired. The computation for determining to which "latent class" to assign a document is relatively straightforward. Winters[21] has proposed a modification in Baker's scheme to make computation more practical. No experimental results have appeared yet for this method, though they are promised by Winters.

Lefkovitz and Prywes[22,23] have proposed a term hierarchy generating process, using what they call inclusive and exclusive partitions. This hierarchy, which can have more than two levels aids in the process of request formulation by steering the requester away from requests specifying co-occurrence of terms which do not co-occur in the document collection. Documents are not explicitly classified in this system.

Doyle[24] has experimented with a multi-level hierarchy generation program written by Ward and Hook. This program operates on indexed documents, first grouping documents whose sets of terms are most nearly alike, then forming larger and larger groups in such an order that at each stage of the process, the most similar remaining groups are combined, and finally the last two groups are joined. This process is in a sense the inverse of successive partitions. Each group at any stage can be thought of as a node in a tree, whose successor nodes are the nodes representing the groups out of which it was formed. Doyle attempts to identify the nodes of this tree with concepts, but with limited success.

Other work on automatic classification is reported in Steven's Statistical Association Methods [24].

V. EVALUATION, POSSIBLE APPLICATIONS, AND RECOMMENDATIONS

The experimental work which has been done in the area of automatic classification has not been every extensive and has been more concerned with comparison of the results of automatic classification versus manual classification than with investigation of how automatically derived classification can facilitate retrieval. Research in generation of two-level hierarchies or placement of documents in them is further along the road to practical application than research in multi-level hierarchies. But a two-level classification system, while an important step toward a system with more than two levels, will probably not have a very great impact on information retrieval systems; what it does can be, and often is, done manually without a great expenditure of effort. A multi-level classification system, in which nodes would represent concepts or terms and in which the tree structure would display relations such as generic-specific, seems to be considerably further from realization than a two-level scheme.

Needham, however, has an interesting basis for an optimistic viewpoint. He says that human classifiers of terms know too much about the meanings of the terms and their relations to each other, and may make a classification system which is more detailed than necessary for a particular set of documents. The computer, having at its disposal only the statistical data for the document set of interest, will not make a classification schedule with unnecessary detail, because it will have no other source of information to draw on. The obvious question is: will this schedule have enough detail to be useful to people?

One problem which lurks in the background of most statistical association techniques is that the number of possible associations of n terms with each other, excluding self-associations, is $\frac{n(n-1)}{2}$. Thus, for large n, the matrix manipulations can become quite unwieldy, even when using special techniques for dealing with sparse matrices.

The following is a list of possible applications of statistical association techniques to Project VECTOR/ROSE:

A. In system establishment

1. Identification of synonyms and near-synonyms, hence thesaurus generation.

2. Use of diagonal elements of the square of the word association matrix, which correspond roughly to how tightly bound a word is to the other words in a document set, in determining what words should be included in the controlled vocabulary or choosing a thesaur from a synonym set; this is an alternative to be compared with Dennis's vocabulary generation approach, etc.

3. Determination of what terms to pre-coordinate or link by scanning of text or abstracts.

4. (possibly) Automatic generation of a classification system for terms or documents.

5. Providing data for man-machine indexing or classification.

6. Providing data for selection of documents to include in the system.

B. In system operation

1. Automatic expansion of a request to include associated terms.

2. Assisting the user in formulation of request by displaying associated terms and the relationships through which they are associated.

REFERENCES

1. Cyril Cleverdon, Jack Mills and Michael Keen, Aslib Cranfield Research Project; "Factors Determining the Performance of Indexing Systems", Vol. 1, 1966.

2. Lauren B. Doyle; "Is Automatic Classification a Reasonable Application of Statistical Analysis of Text?", JACM, Vol.12, p. 473.

3. M. E. Maron and J. L. Kuhns; "On Relevance, Probabilistic Indexing and Information Retrieval", JACM, Vol. 7, 1960, p. 216.

4. H. P. Edmundson and R. E. Wyllys; "Automatic Abstracting and Indexing -- Survey and Recommendations", CACM, Vol. 4., 1961, p. 226.

5. Fred. J. Damerau; "An Experiment in Automatic Indexing", American Documentation, Vol. 16, p. 283, 1965.

6. John O'Connor; "Automatic Subject Recognition in Scientific Papers: An Empirical Study", JACM, Vol. 12, p. 490, 1965.

7. John O'Connor; "Mechanized Indexing Methods and Their Testing", JACM, Vol. 11, p. 437, 1964.

8. Sally F. Dennis; "The Construction of a Thesaurus Automatically From a Sample of Text", in Statistical Association Methods for Mechanized Documentation, ed. by Mary E. Stevens, et al., p. 61, 1965.

9. H. Edmund Stiles; "The Association Factory in Information Retrieval", JACM, Vol. 8, p. 271, 1961.

10. Lauren B. Doyle; "Semantic Road Maps for Literature Searchers", JACM, Vol. 8, p. 553, 1961.

11. Lauren B. Doyle; "Indexing and Abstracting by Association", American Documentation, Vol. 13, p. 378, 1962.

12. Vincent E. Guiliano and Paul E. Jones; "Linear Associative Information Retrieval", in Vistas in Information Handling, Vol. I, ed. by Paul W. Howerton, p. 30, 1963.

13. Gerard Salton; "A Document Retrieval System for Man-Machine Interaction", ACM Proceedings of the 19th National Conference, 1964, p. L2.3-1.

14. Joseph Becker and Robert M. Hayes; "Information Storage and Retrieval", Wiley, 1963, Chapt. 14.

15. Melvin E. Maron; "Automatic Indexing; An Experimental Inquiry", JACM, Vol. 8, p. 404, 1961.

16. J. H. Williams; "Results of Classifying Documents with Multiple Discriminant Functions", Steven's Stat. Assoc. Meth. p. 217.

17. R. Needham; "A Method for Using Computers in Information Classification", IFIP 1962, p. 284.

18. Needham and K. Sparck Jones; "Keywords and Clumps", Jour. of Doc. Vol. 20, p. 5.

19. Borko and Bernick; "Automatic Document Classification", JACM Vol. 10, p. 151; Vol. 11, p. 138.

20. F. Baker; "Latent Class Analysis as an Association Model for Information Retrieval", Steven's Stat. Assoc. Meth., p. 149.

21. Wm. K. Winters; "A Modified Method of Latent Class Analysis", JACM, Vol. 12, p. 356.

22. D. Lefkovitz and N. S. Prywes; "Automatic Stratification of Information", AFIPS Conference Proceedings, Vol. 23, 1963 SJCC, p. 229.

23. D. Lefkovitz; "The Application of the Digital Computer to the Problem of a Document Classification System", in Colloquium on Technical Preconditions for Retr. Cent. Op., ed. by Cheydleur, p. 133.

24. Lauren B. Doyle; "Some Comprises Between Word Grouping and Document Grouping", in Steven's Statistical Assoc. Methods, p. 15, See also ref. 2 of this report.

25. Lauren B. Doyle; "Expanding the Editing Function in Language Data Processing", CACM, Vol. 9, April 1965, p. 238.

26. V. Guiliano; "Postscript...", in Stevens Stat. Assoc. Meth., p. 259.